Phylogenetic positions of '*Candidatus* Phytoplasma asteris' and *Spiroplasma kunkelii* as inferred from multiple sets of concatenated core housekeeping proteins

Yan Zhao, Robert E. Davis and Ing-Ming Lee

Molecular Plant Pathology Laboratory, USDA-Agriculture Research Service, BARC-West, 10300 Baltimore Avenue, Beltsville, MD 20705, USA

Correspondence
Yan Zhao
zhaoy@ba.ars.usda.gov

Phytopathogenic mollicutes, which include spiroplasmas and phytoplasmas, are cell wall-less bacteria that parasitize plant hosts and insect vectors. Knowledge of the evolution of these agents is important in understanding their biology. The availability of the first complete phytoplasma and several partial spiroplasma and phytoplasma genome sequences made possible an investigation of evolutionary relationships between phytopathogenic mollicutes and other micro-organisms, especially Gram-positive bacteria, using a comparative genomics approach. Genome data from a total of 41 bacterial species were used in the analysis. Sixty-one conserved proteins were selected from each species for the construction of a hypothetical phylogenetic tree. The genes encoding these selected proteins are among a core of genetic elements that constitute a hypothetical minimal genome. The proteins were concatenated into five superproteins according to their functional categories, and phylogenetic trees were reconstructed using distance, parsimony and likelihood methods. Phylogenetic trees based on the five sets of concatenated proteins were congruent in both clade topology and relative branching length. *Spiroplasma kunkelii* and phytoplasmas clustered together with other mollicutes, forming a monophyletic group. Phytoplasmas diverged from spiroplasmas and mycoplasmas at early stages in the evolution of mollicutes. Branch lengths on the phylogenetic trees were noticeably longer in the *Mollicutes* clade, suggesting that the genes encoding the five sets of proteins evolved at a greater rate in this clade than in other clades. This observation reinforces the concept that mollicutes have rapidly evolving genomes.

INTRODUCTION

Phytopathogenic spiroplasmas and phytoplasmas are small, cell wall-less bacteria that cause disease in more than 300 vegetable, ornamental and perennial species, representing over 100 plant families (Bové, 1997; Davis *et al.*, 1972; Lee *et al.*, 2000). These phytopathogenic agents are restricted, in plants, to sieve cells of phloem tissue and are transmitted from diseased to healthy plants by insect vectors, mainly leafhoppers and psyllids. Spiroplasmas and phytoplasmas belong to the class *Mollicutes* and have been thought to be derived from ancestral low-G+C Gram-positive bacteria, possibly some ancient members of the *Bacillus–Clostridium* group, through retrogressive evolution and massive genome reduction (Razin *et al.*, 1998). Since phytopathogenic mollicutes parasitize both plant hosts and insect vectors, knowledge of the evolution of these agents could be helpful in understanding their biology.

Genetic relatedness or evolutionary relationships among micro-organisms, including mollicutes, have often been proposed based on analyses of 16S rRNA genes (Gasparich *et al.*, 2004; Gundersen *et al.*, 1994; Olsen & Woese, 1993; Stackebrandt & Goebel, 1994; Stackebrandt *et al.*, 1997; Weisburg *et al.*, 1989; Woese, 1987), mainly because 16S rRNA gene sequences are ubiquitous and well-conserved across the spectrum of prokaryotes, and are known for many bacterial species. However, the highly conserved nature of the 16S rRNA gene tends to limit its power to resolve closely related organisms that diverged at almost the same time (Fox *et al.*, 1992; Woese, 1987). Individual, conserved protein-encoding genes have also been analysed to build phylogenetic trees of prokaryotes. Nevertheless, the topologies of such protein-based phylogenetic trees do not always agree with that of a 16S rRNA gene tree (Brown & Doolittle, 1997; Golding & Gupta, 1995). This can be attributed to either extensive lateral gene transfer or degradation of

Published online ahead of print on 10 June 2005 as DOI 10.1099/ijs.0.63655-0.

Abbreviations: ML, maximum likelihood; MP, maximum parsimony; NJ, neighbour joining.

phylogenetic signals caused by saturation of amino acid substitutions during the evolution of organisms (Brochier *et al.*, 2000; Brown *et al.*, 2001; Doolittle, 1999; Forterre & Philippe, 1999). The recent availability of multiple complete genome sequences from diverse bacterial taxa creates unprecedented opportunities to explore new phylogenetic approaches based on comparative analysis of full gene complements or large subsets thereof (Brown *et al.*, 2001; Daubin *et al.*, 2002; Wolf *et al.*, 2001). One such new approach involves the use of large combined or concatenated orthologous proteins to construct a universal tree; the concatenated protein data lead to significant amplification of phylogenetic signals and increased resolving power (Brown *et al.*, 2001; Wolf *et al.*, 2001). Another novel

Table 1. List of bacterial genomes used in this study

Species name	Strain	Abbreviation	GenBank accession no.*
'*Candidatus* Phytoplasma asteris'-related	Aster yellows WB	Astyel	NA[a]
Bacillus anthracis	Ames	Bacant	AE016879
Bacillus cereus	ATCC 14579^T	Baccer	AE016877
Bacillus halodurans	C-125	Bachal	BA000004
Bacillus subtilis subsp. *subtilis*	168	Bacsub	AL009126
Clostridium acetobutylicum	ATCC 824^T	Cloace	AE001437
Clostridium perfringens	13	Cloper	BA000016
Corynebacterium diphtheriae	NCTC 13129	Cordip	BX248353
Corynebacterium efficiens	YS-314^T	Coreff	BA000035
Corynebacterium glutamicum	ATCC 13032^T	Corglu	BA000036
Enterococcus faecalis	V583	Entfae	AE016830
Lactobacillus johnsonii	NCC 533	Lacjoh	AE017198
Lactobacillus plantarum	WCFS1	Lacpla	AL935263
Lactococcus lactis subsp. *lactis*	Il1403	Laclac	AE005176
Listeria innocua	Clip11262	Lisinn	AL592022
Listeria monocytogenes	EGD-e	Lismon	AL591824
Mesoplasma florum	L1^T	Mesflo	AE017263
Mycobacterium bovis subsp. *bovis*	AF2122/97	Mycbov	BX248333
Mycobacterium leprae	TN	Myclep	AL450380
Mycobacterium tuberculosis	CDC1551	Myctub	AE000516
Mycoplasma gallisepticum	R	Mycgal	AE015450
Mycoplasma genitalium	G-37^T	Mycgen	L43967
Mycoplasma mobile	163K^T	Mycmob	AE017308
Mycoplasma mycoides subsp. *mycoides* SC	PG1	Mycmyc	BX293980
Mycoplasma penetrans	HF-2	Mycpen	BA000026
Mycoplasma pneumoniae	M129	Mycpne	U00089
Mycoplasma pulmonis	UAB CTIP	Mycpul	AL445566
Nostoc sp.	PCC 7120	Nossp	BA000019
Oceanobacillus iheyensis	HTE831^T	Oceihe	BA000028
'*Candidatus* Phytoplasma asteris'-related	Onion yellows M	Oniyel	AP006628
Spiroplasma kunkelii	CR2-3x	Spikun	NC_003999[b]
Staphylococcus aureus subsp. *aureus*	MW2	Staaur	BA000033
Staphylococcus epidermidis	ATCC 12228	Staepi	AE015929
Streptococcus agalactiae	2603V/R	Straga	AE009948
Streptococcus pneumoniae	TIGR4	Strpne	AE005672
Streptococcus pyogenes	MGAS8232	Strpyo	AE009949
Streptomyces avermitilis	MA-4680^T	Strave	BA000030
Streptomyces coelicolor	A3(2)	Strcoe	AL645882
Synechocystis sp.	PCC 6803	Synsp	BA000022
Thermoanaerobacter tengcongensis	MB4^T	Theten	AE008691
Ureaplasma parvum serovar 3	ATCC 700970	Urepar	AF222894

SC, small colony.

*Data were accessed at: *a*, http://www.oardc.ohio-state.edu/phytoplasma/genome.htm and *b*, http://www.genome.ou.edu/spiro.html; NA, not available.

approach involves comparison of topologies of individual gene trees and focuses on congruence of tree topologies, which permits identification of a core of genes that share a common history and have undergone fewer lateral transfers (Daubin *et al.*, 2002). In the present study, by combining the above two approaches, we constructed a hypothetical phylogenetic tree to propose the evolutionary positions of phytopathogenic mollicutes. Analyses of 61 core housekeeping proteins from 41 bacterial species suggested a consensus phylogeny: phytopathogenic spiroplasmas and phytoplasmas clustered with other mollicutes, forming a monophyletic clade. Phytoplasmas diverged from spiroplasmas and mycoplasmas at an early stage in the evolution of the class *Mollicutes*.

METHODS

Source of sequence data. The amino acid sequences of proteins that are encoded in the completely sequenced genomes (Table 1) were extracted from the annotated genome data deposited in GenBank, DDBJ and EMBL. Incomplete genome sequences of *Spiroplasma kunkelii* strain CR2-3x and '*Candidatus* Phytoplasma asteris'-related strain AY-WB were retrieved from the University of Oklahoma's Advanced Center for Genome Technology Internet web site at http://www.genome.ou.edu/spiro.html and the Ohio State University's phytoplasma genome sequencing web site at http://www.oardc.ohio-state.edu/phytoplasma/genome.htm respectively, as assembled contigs. The individual contigs were analysed by using the heuristic models of GeneMark and GeneMark.hmm programs (Besemer & Borodovsky, 1999) for identifying potential open reading frames. The predicted protein-encoding genes were annotated following a search of the National Center for Biotechnology Information (NCBI)'s non-redundant protein database using the position-specific iterated and pattern-hit iterated BLAST programs (Altschul *et al.*, 1997), and a search of the Clusters of Orthologous Groups (COG) database using the program COGnitor (Tatusov *et al.*, 2001).

Selection and concatenation of datasets. A complete set of ribosomal proteins (51), DNA polymerase III subunits (5), DNA-dependent RNA polymerase subunits (4), excision nuclease subunits (3), glycolysis enzymes (9) and components of Sec-dependent secretion machinery (6) were extracted from genome sequence data of '*Candidatus* Phytoplasma asteris'-related strain OY-M (associated with onion yellows) (Oshima *et al.*, 2004). This set of 78 proteins was used as queries for BLAST2 searches against completely sequenced bacterial genomes, listed in Table 1, at the DDBJ's Genome Information Broker web site (http://gib.genes.nig.ac.jp/). Among the 78 proteins, 61 are ubiquitously present in all the genomes and are well conserved. These 61 proteins are also present in the partially sequenced genomes of '*Candidatus* Phytoplasma asteris'-related strain AY-WB and *Spiroplasma kunkelii*, and therefore were selected for phylogenetic analysis. Orthologues of these selected proteins were retrieved from the 41 genomes, divided into five subsets according to their functional categories and concatenated head-to-tail (Table 2).

Phylogenetic analyses. The amino acid sequences of the selected proteins were compiled and concatenated in FASTA format. The concatenated sequences were first aligned using CLUSTAL_X (version 1.81) by selecting the 'do complete alignment' option with default parameters (Jeanmougin *et al.*, 1998; Thompson *et al.*, 1997). Each output alignment was converted to NBRF/Pir format and was trimmed using GBLOCKS (version 0.91b) to eliminate poorly aligned positions (Castresana, 2000). The trimmed alignment was converted to MEGA or PHYLIP format for phylogenetic analyses. Distance analyses were performed with the MEGA2 (version 2.1; Kumar *et al.*, 2001) or CLUSTAL_X (version 1.81; Thompson *et al.*, 1997) package using the neighbour-joining (NJ) method with Gamma-distance model for multiple substitutions at the same amino acid site. Parsimony analyses were conducted with PHYLIP (version 3.62; Felsenstein, 1989, 2004) using the Protein Parsimony algorithm (ProtPars). Maximum-likelihood (ML) analyses were performed with PHYLIP (version 3.62) using the Protein Maximum Likelihood program (ProtML) (Felsenstein & Churchill, 1996) with the Jones–Taylor–Thornton probability model option (Jones *et al.*, 1992). The reliability of each phylogenetic analysis was subjected to a bootstrap test with 1000 replicates. The five best trees from the default output of each analysis were used as input for the CONSENSE program of the PHYLIP (version 3.62) package to generate a consensus tree. Phylogenetic trees were viewed using TreeExplorer of the MEGA2 package (Kumar *et al.*, 2001) and the PhyloDraw program (Choi *et al.*, 2000).

RESULTS AND DISCUSSION

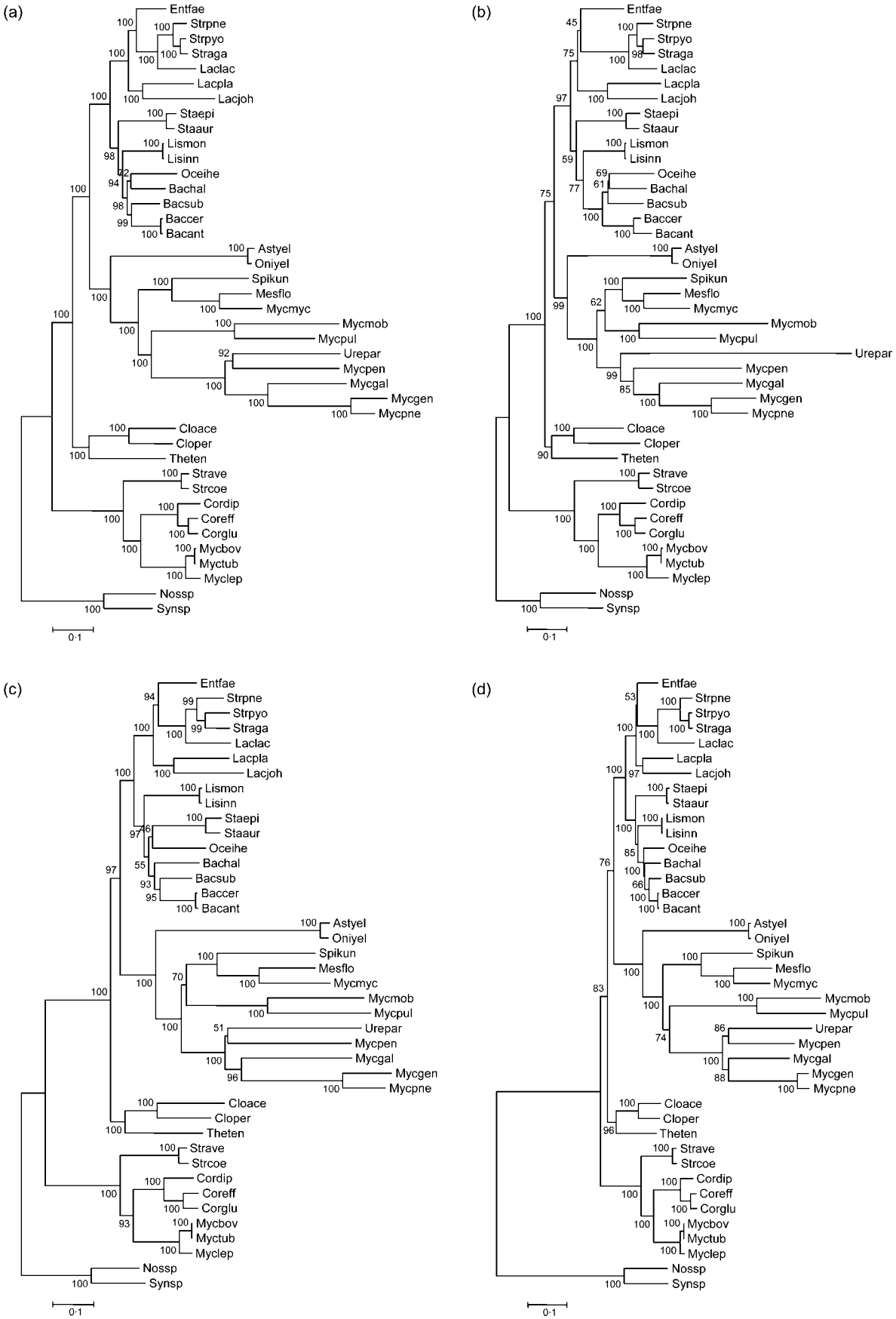
Phytopathogenic spiroplasmas and phytoplasmas are among the micro-organisms that possess a small genome with a gene set approaching the minimal complement necessary for cellular life and pathogenesis. Reconstruction

Table 2. List of concatenated proteins

Protein set	Concatenated components*	Alignment column	Trimmed column
Ribosomal proteins	RpsJ-RplC-RplD-RplW-RplB-RpsS-RplV-RpsC-RplP-RpmC-RpsQ-RplN-RplX-RplE-RpsN-RpsH-RplF-RplR-RpsE-RplO-RplA-RplK-RplM-RpsB-RpsD-RpsG-RpsI-RpsK-RpsL-RpsM-RpsO-RplQ-RplS-RplT-RplU-RpmA-RpmE-RpmF-RpmI-RpmJ-RpsF-RpsP-RpsR-RpsT	8622	4117
Glycolytic enzymes	Pfk-Fba-Pgk-Eno-Pyk†	2798	1141
DNA replication/repair proteins	UvrA-UvrB-UvrC-DnaE (PolC)	4669	1810
RNA polymerase subunits	Alpha-Beta-Beta prime-Delta	3770	1747
Protein secretion components	SecA-SecY-SRP54-SRPR	3429	1345

*Protein names were adopted from Clusters of Orthologous Groups (http://www.ncbi.nlm.nih.gov/COG/) and Kyoto Encyclopedia of Genes and Genomes Pathway database (http://www.genome.jp/kegg/metabolism.html).

†Eno, enolase; Fba, fructose-bisphosphate aldolase; Pfk, 6-phosphofructokinase; Pgk, phosphoglycerate kinase; Pyk, pyruvate kinase.


(a)
Entfae
Strpne
Strpyo
Straga
Laclac
Lacpla
Lacjoh
Staepi
Staaur
Lismon
Lisinn
Oceihe
Bachal
Bacsub
Baccer
Bacant
Astyel
Oniyel
Spikun
Mesflo
Mycmyc
Mycmob
Mycpul
Urepar
Mycpen
Mycgal
Mycgen
Mycpne
Cloace
Cloper
Theten
Strave
Strcoe
Cordip
Coreff
Corglu
Mycbov
Myctub
Myclep
Nossp
Synsp
0·1
(b)
Entfae
Strpne
Strpyo
Straga
Laclac
Lacpla
Lacjoh
Staepi
Staaur
Lismon
Lisinn
Oceihe
Bachal
Bacsub
Baccer
Bacant
Astyel
Oniyel
Spikun
Mesflo
Mycmyc
Mycmob
Mycpul
Urepar
Mycpen
Mycgal
Mycgen
Mycpne
Cloace
Cloper
Theten
Strave
Strcoe
Cordip
Coreff
Corglu
Mycbov
Myctub
Myclep
Nossp
Synsp
0·1
(c)
Entfae
Strpne
Strpyo
Straga
Laclac
Lacpla
Lacjoh
Lismon
Lisinn
Staepi
Staaur
Oceihe
Bachal
Bacsub
Baccer
Bacant
Astyel
Oniyel
Spikun
Mesflo
Mycmyc
Mycmob
Mycpul
Urepar
Mycpen
Mycgal
Mycgen
Mycpne
Cloace
Cloper
Theten
Strave
Strcoe
Cordip
Coreff
Corglu
Mycbov
Myctub
Myclep
Nossp
Synsp
0·1
(d)
Entfae
Strpne
Strpyo
Straga
Laclac
Lacpla
Lacjoh
Staepi
Staaur
Lismon
Lisinn
Oceihe
Bachal
Bacsub
Baccer
Bacant
Astyel
Oniyel
Spikun
Mesflo
Mycmyc
Mycmob
Mycpul
Urepar
Mycpen
Mycgal
Mycgen
Mycpne
Cloace
Cloper
Theten
Strave
Strcoe
Cordip
Coreff
Corglu
Mycbov
Myctub
Myclep
Nossp
Synsp
0·1

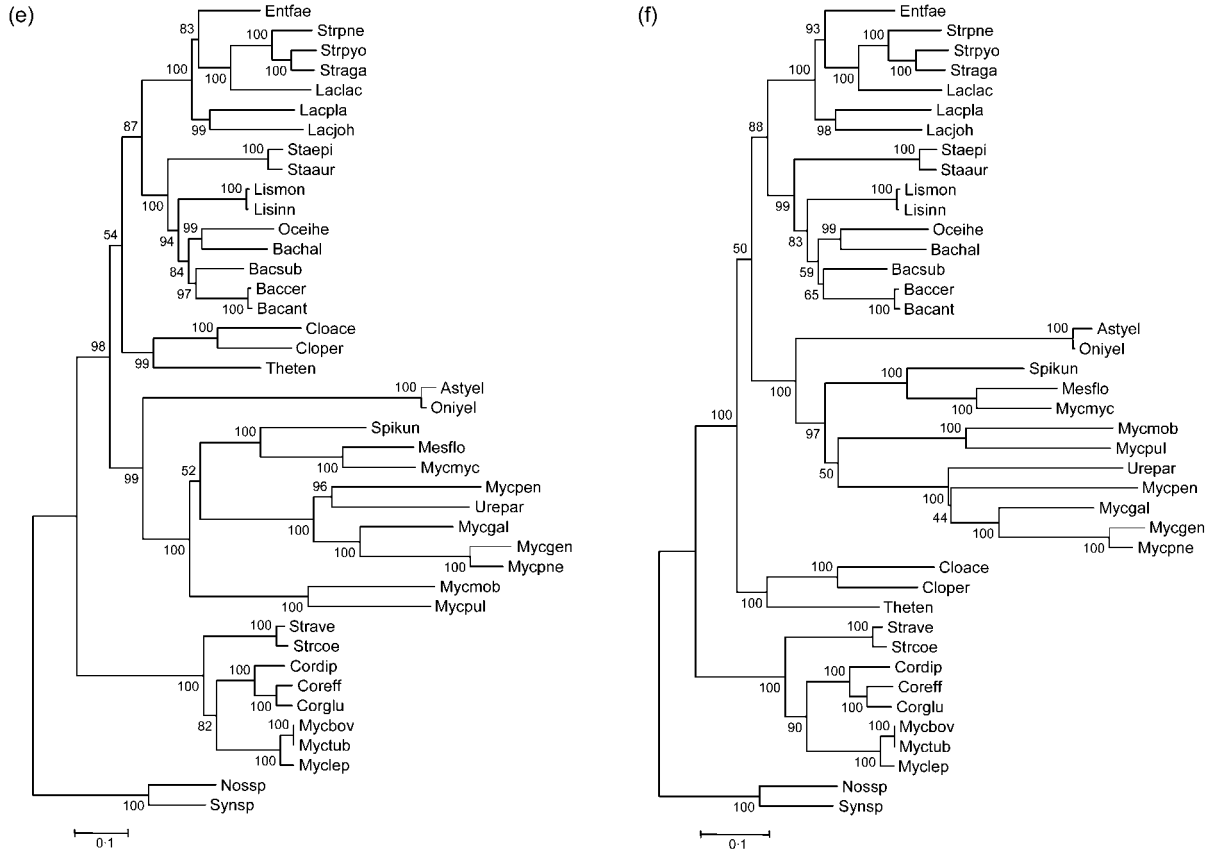


Fig. 1. Phylogenetic trees derived from distance analysis of five sets of concatenated proteins. The sequence sets were analysed using the NJ algorithm (MEGA 2.1) with the Gamma parameter set at 2·5. The reliability of each tree topology was subjected to a bootstrap test. Numbers at nodes indicate bootstrap support values as a percentage of 1000 replications. Rooting with the outgroup (*Nostoc* sp. and *Synechocystis* sp.) in each tree was forced. (a) Tree topology based on 44 ribosomal proteins; (b) tree topology based on five glycolytic enzymes; (c) tree topology based on four DNA replication/repair proteins; (d) tree topology based on four RNA polymerase subunits; (e) tree topology based on four protein secretion components; and (f) tree topology based on SecA and SecY. See Table 1 for a key to abbreviated bacterial names.

of the evolutionary history of these unique pathogenic agents could help us to understand how such minuscule cell wall-less bacteria might have evolved to adapt to their physiological niches and gained pathogenicity while undergoing massive genome reduction. We took a protein tree-based approach to study the evolutionary relationships among bacterial species. The 61 core housekeeping proteins selected for phylogenetic analysis are evolutionarily conserved and are involved in fundamental life processes such as DNA replication and repair, RNA transcription, protein synthesis, protein translocation and carbohydrate metabolism. Congruency analysis of phylogenetic trees based on the five groups of concatenated proteins enabled us to present a consensus phylogeny with special emphasis on mollicutes.

Ribosomal protein tree

Ribosomal proteins are key protein components of the ribosome, a ubiquitous cellular apparatus that translates genetic information encoded in mRNA into proteins. Ribosomal proteins are conserved among bacterial species not only at individual gene/protein sequence levels but also at the gene organization level. Most ribosomal protein genes are clustered in several highly conserved operons, such as the S10-spc-alpha superoperon, which ensures coordinated expression of the genes. Although duplication and possible horizontal transfer of individual ribosomal protein genes have been described in some recent reports (Brochier *et al.*, 2000; Makarova *et al.*, 2001), such events are considered infrequent. Large-scale comparative genomics studies suggest that ribosomal protein genes are among a core of genes that share a common history (Daubin *et al.*, 2002) and carry a strong phylogenetic signal (Wolf *et al.*, 2001). The 44 ribosomal proteins used in the present study were concatenated head-to-tail and treated as a single protein sequence. The initial alignment of the concatenated protein sequences from 41 species contained 8622 columns. The removal of gaps and poorly aligned regions that might not be homologous or might have been saturated by multiple substitutions (Castresana, 2000) resulted in a final

alignment of 4117 columns. Use of concatenated multiple ribosomal protein sequences allowed for amplification of phylogenetic signals and reduction of potential noise caused by possible lateral gene transfer events. Phylogenetic analysis using a distance method (NJ algorithm) resulted in the tree topology shown in Fig. 1(a). The tree topology was supported by high bootstrap values. Phylogenetic analyses using maximum-parsimony (MP) and ML algorithms generated ribosomal protein trees with nearly identical topologies (data not shown; see consensus phylogeny below).

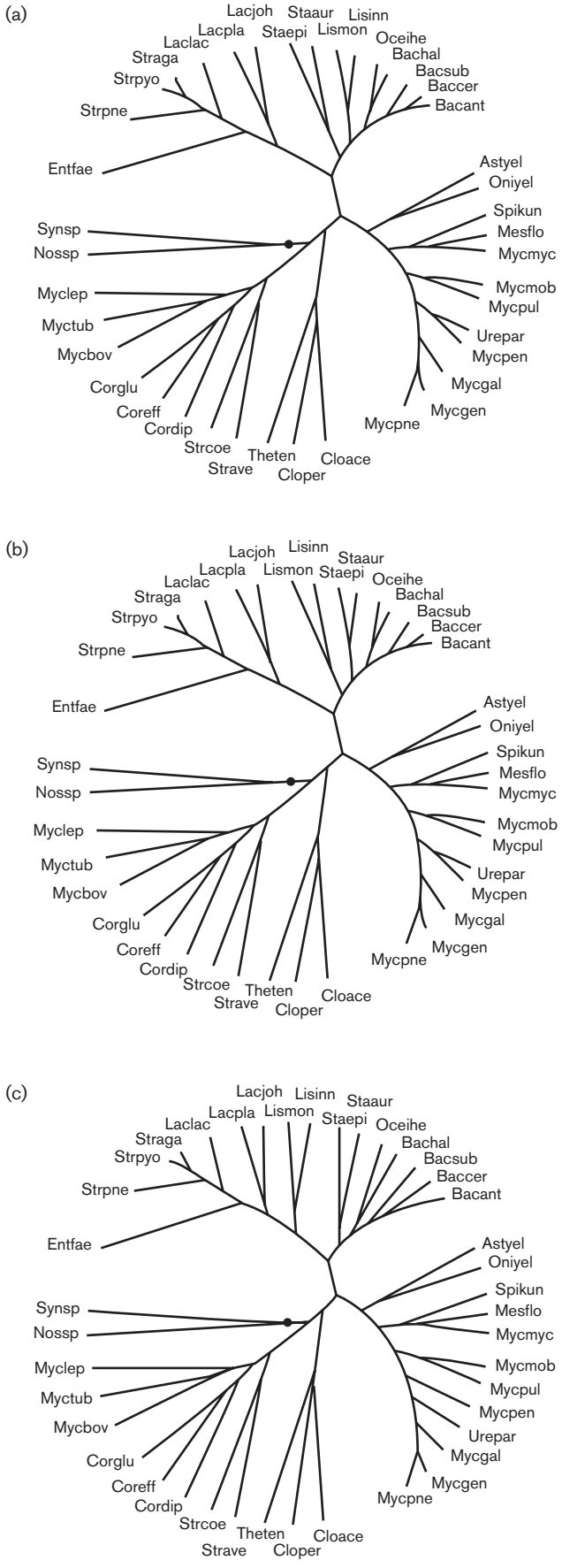
In the phylogeny inferred from the combined ribosomal protein data, phytopathogenic mollicutes clustered together with animal- and human-pathogenic mollicutes as well as non-pathogenic members of the class *Mollicutes*, forming a monophyletic group. It appeared that phytoplasmas diverged from spiroplasmas and mycoplasmas at an early stage in the evolution of mollicutes. The *Mollicutes* clade appeared to be paraphyletic to a clade consisting of *Bacillales* and '*Lactobacillales*', suggesting that the *Mollicutes* clade and the *Bacillales*–'*Lactobacillales*' clade share a common *Clostridium*-like ancestor. The branch lengths tended to be longer in the *Mollicutes* clade, suggesting that the genes encoding the ribosomal proteins evolved at a greater rate in this clade than in other clades.

Congruence of the ribosomal protein tree with other protein trees

The other protein datasets used in this study included five key enzymes that are involved in glycolysis (Table 2). Glycolysis is a universal metabolic pathway that converts glucose into pyruvate with the concomitant production of a relatively small amount of ATP. This pathway probably developed before there was sufficient oxygen in the atmosphere to sustain more-effective methods of energy extraction (Barnett, 2003). When aerobic organisms evolved, a more-efficient energy-harvesting pathway, namely the tricarboxylic acid (TCA) cycle, developed; oxidative phosphorylation steps were added onto glycolysis (Barnett, 2003). Glycolytic enzymes are among the most highly conserved enzymes known. It was estimated that one of the key glycolytic enzymes, phosphoglycerate kinase (Pgk), has an evolutionary history of 40 million years (Ciccarese *et al.*, 1989) and has been evolving at a linear rate of 4·8 accepted point mutations per 100 million years (Fothergill-Gilmore, 1986). The value of glycolysis enzymes as a phylogenetic marker has been extensively evaluated (Canback *et al.*, 2002; Chattopadhyay & Chakrabarti, 2003). Recently, Pgk was used to reconstruct a phylogeny of '*Firmicutes*' with special reference to *Mycoplasma* (Wolf *et al.*, 2004). The role of glycolysis in mollicutes is especially critical, as these organisms lack a TCA cycle (Pollack *et al.*, 1997). In the present study, the initial alignment and the trimmed alignment of the concatenated glycolytic enzymes contained 2798 and 1141 columns, respectively. Phylogenetic analysis of the trimmed alignment generated a phylogeny (Fig. 1b) with a tree topology that was almost identical to that inferred from combined ribosomal proteins. The only difference was that, in the ribosomal protein tree (Fig. 1a), the members of the hominis group (represented by *Mycoplasma pulmonis* and *Mycoplasma mobile*) formed a sister lineage with members of the pneumoniae group, whereas in the glycolytic enzyme tree (Fig. 1b), the hominis group formed a sister lineage with members of the mycoides group, although the bootstrap value for this sister lineage relationship was lower than that for the hominis group–pneumoniae group sister lineage relationship in the ribosomal protein tree.

DNA polymerase III alpha-subunit, excinuclease ABC subunits and DNA-dependent RNA polymerase subunits were also included in our datasets. These proteins are among the core components whose genes share a common history (Daubin *et al.*, 2002) and are involved in genetic information storage, repair and transcription. They carry strong phylogenetic signals and have greater resolving power than 16S rRNA gene sequences (Klenk & Zillig, 1994; Mollet *et al.*, 1997; Teeling *et al.*, 2004). A phylogenetic tree constructed based on concatenated DNA polymerase III subunit alpha (DNApol) and excinuclease subunits A, B and C (UVR) exhibited a topology that was almost identical to that of the ribosomal protein tree (Fig. 1a), with only minor differences at a low taxonomic level within the *Bacillales* group (Fig. 1c) where the positions of *Listeria* species and *Staphylococcus* species were switched, and *Mycoplasma pulmonis* and *Mycoplasma mobile* were grouped with the mycoides clade. Furthermore, in the UVR–DNApol tree, *Oceanobacillus iheyensis* was clustered with *Staphylococcus* species and paraphyletic to *Bacillus halodurans*, whereas in the ribosomal protein tree *O. iheyensis* was clustered with *B. halodurans* and was paraphyletic to *Staphylococcus* species. The phylogeny inferred from the four combined DNA-dependent RNA polymerase subunits gave a tree topology (Fig. 1d) that was almost exactly the same as that of the ribosomal protein tree.

In bacteria, the secretion of proteins from the site of their synthesis to outside the cytoplasmic membrane is mediated by multiple protein translocation systems, among which the Sec-dependent protein translocation machinery is the most prominent and is ubiquitous to all bacteria. The Sec machinery consists of a heterotrimeric transmembrane channel, SecYEG, and a peripheral homodimeric ATPase, SecA (den Blaauwen & Driessen, 1996). Highly hydrophobic preproteins that are translocated via the Sec machinery may bypass the SecA component and use the signal recognition particle (SRP) pathway to approach the SecYEG channel (Tjalsma *et al.*, 2000). Sequences of both Sec machinery components and SRP components are well conserved among bacteria. In the present study, SecA, SecY, SRP54 and SRPR sequences were selected for phylogenetic analysis. The phylogeny inferred from this set of combined data was slightly different from that suggested by the ribosomal protein data. In this protein translocation tree (Fig. 1e), the *Clostridium* clade clustered together with the



Bacillales–'*Lactobacillales*' clade to form a monophyletic group, paraphyletic to the *Mollicutes* clade. However, when the SRP54 and SRPR sequences were removed from the alignment, the SecA–SecY phylogeny obtained (Fig. 1f) was in excellent agreement with the ribosomal protein phylogeny, although the bootstrap support for some branches was weak. It would be interesting to know whether or not lineage-specific lateral SRP54/SRPR gene transfer occurred after the divergence of the *Clostridium* group and the *Bacillales*–'*Lactobacillales*' group. Since genes encoding SRP54 and SRPR proteins are single copy genes in all bacterial species used in this study, additional orthology information may be helpful in resolving the topology difference caused by the SRP54 and SRPR proteins (Philippe & Forterre, 1999).

A consensus phylogeny

To resolve the minor differences that were present in the phylogenetic trees derived from the five sets of concatenated proteins, two approaches were taken towards building a consensus phylogeny. In the first approach, five output tree files resulting from each phylogenetic analysis method (NJ, MP or ML) were used as input files for computing consensus trees according to the majority rules set by the CONSENSE program of the PHYLIP software suite (Felsenstein, 1989). As shown in Fig. 2(a–c), the consensus trees based on the data from three different phylogenetic analysis methods are in close mutual agreement. The second approach involved further concatenation of the five sets of concatenated proteins into a single superprotein set and subsequent phylogenetic analysis of the superprotein set. The NJ phylogeny inferred by the superprotein (Fig. 3) was identical to the consensus NJ phylogeny computed using the CONSENSE program (Fig. 2a). In the consensus phylogeny, all mollicutes, including plant-pathogenic spiroplasmas and phytoplasmas, appear to be monophyletic and to share a common *Clostridium*-like ancestor. Within the *Mollicutes* clade, phytoplasmas diverged from the rest of the *Mollicutes* at an early stage in the evolution of mollicutes. The divergence between phytoplasmas and the other mollicutes is also indicated by the differences between the genetic code systems used by the phytoplasmas and the rest of the

Fig. 2. Consensus tree topologies based on phylogenetic congruence among five sets of concatenated proteins. Sequence sets were analysed with distance (NJ), MP or ML methods; the resulting output data were used as input for the CONSENSE program of the PHYLIP (version 3.62) package and a consensus topology was generated according to the majority rule. Rooting with the outgroup (*Nostoc* sp. and *Synechocystis* sp.; indicated by a filled circle at the node) in each consensus tree was forced. (a) NJ consensus tree; the analysis was performed using CLUSTAL_X (version 1.81) with Kimura's Gamma model; (b) MP consensus tree; the analysis was performed using PHYLIP with the ProtPars algorithm; and (c) ML consensus tree; the analysis was performed using PHYLIP with the ProtML algorithm. See Table 1 for a key to abbreviated bacterial names.

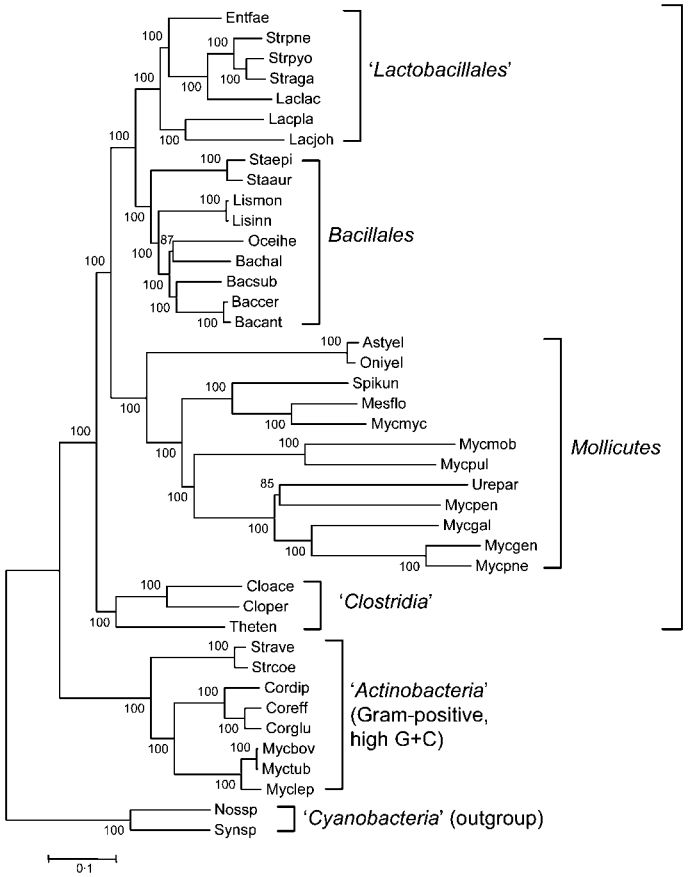


Fig. 3. Phylogenetic tree constructed from combined protein data. Sequences of 61 proteins, from all five selected datasets, were concatenated into a single super-protein set. The combined sequence set was analysed using the NJ algorithm (MEGA 2.1) with the Gamma parameter set at 2·5. A bootstrap test of 1000 replications was performed to examine the reliability of the phylogeny. Numbers at nodes indicate percentage bootstrap support values. Rooting with the outgroup in the tree was forced. See Table 1 for a key to abbreviated bacterial names.

mollicutes examined in this study. Whereas phytoplasmas and acholeplasmas use the standard genetic code system, spiroplasmas and other mollicutes use an uncommon genetic code system, in which the triplet TGA encodes a tryptophan rather than being a translation termination codon (Citti *et al.*, 1992). The plant-pathogenic species *Spiroplasma kunkelii*, together with its non-helical siblings, which include the animal-pathogenic species *Mycoplasma mycoides* and the non-pathogenic species *Mesoplasma florum*, formed the mycoides clade. The mycoides clade also branched off relatively early from the other human- and animal-pathogenic mollicutes. The remaining mollicutes were clustered into two major groups, namely the hominis group and the pneumoniae group. This multiple-protein consensus phylogeny coincided well with the phylogeny inferred from ribosomal 16S rRNA genes (Fig. 4). However, one problem with all the phylogenetic trees in this study is that they are all based on limited available sequences and mostly originate from taxa that are only distantly related (except the two '*Candidatus* Phytoplasma asteris'-related species). Thus, there is a risk that long-branch attraction can result in topologies that do not reflect the true phylogenies. Because of the lack of more-extensive sequence data, it is difficult to perform a more-extensive study at this time. Nevertheless, the multiple-protein sequence-based tree has branch lengths that are almost twice as long as those of the 16S rRNA gene tree. Therefore, the multiple-protein approach appears to offer a greater resolving power than the 16S rRNA gene approach, and should be helpful in determining phylogenetic positions of closely related species.

In the consensus phylogeny, the branch lengths are notably longer in the *Mollicutes* clade than in other clades. This is not surprising since the same pattern was seen in each of the five phylogenies reconstituted from concatenated proteins. The presence of relatively long branches suggests that the proteins evolved at a greater rate in the *Mollicutes* clade than in other clades. The rapid rate of evolution of these core housekeeping protein genes may reflect rapid evolutionary change in whole mollicute genomes.

Conclusion

The class *Mollicutes* consists of a group of genetically heterogeneous wall-less prokaryotes. Until recently, their taxonomy has been based on biochemical and phenotypic criteria (Razin *et al.*, 1998) that are inaccessible for uncultured prokaryotes. The molecular technology developed during the past two decades has advanced the systematics of mollicutes, especially uncultured plant-pathogenic phytoplasmas. Phylogenetic analyses based on conserved 16S rRNA gene sequences made it possible to envision the molecular genetic relatedness among diverse mollicutes and their evolutionary relationships with walled prokaryotes

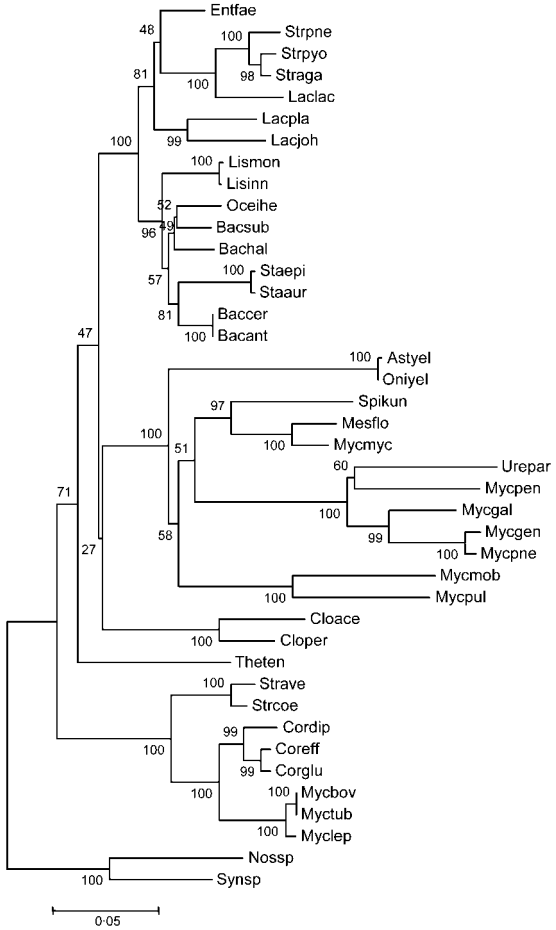


Fig. 4. Phylogenetic tree constructed from 16S rRNA gene sequences. Full-length 16S rRNA gene sequences from 41 bacterial species were aligned using CLUSTAL_X (version 1.81). The alignment was trimmed using GBLOCKS (version 0.91b) to remove poorly aligned positions. The trimmed alignment was analysed using the NJ algorithm (MEGA 2.1) with a bootstrap test of 1000 replications. Numbers at nodes indicate percentage bootstrap support values. Rooting with the outgroup (*Nostoc* sp. and *Synechocystis* sp.) in the tree was forced. See Table 1 for a key to abbreviated bacterial names.

(Gasparich *et al.*, 2004; Gundersen *et al.*, 1994; Lee *et al.*, 2000; Weisburg *et al.*, 1989; Woese, 1987). Several members of the class *Mollicutes* have been reclassified based on phylogenies inferred by 16S rRNA gene sequence analyses. To date, the class *Mollicutes* consists of four orders and eight formal genera. Moreover, 16S rRNA gene-based phylogenies have begun to reveal the evolutionary history of mollicutes; although they loosely formed a monophyletic group, mollicutes appeared to be comprised of several phylogenetic clusters that were deeply divergent from one another, as indicated by long branch lengths on phylogenetic trees. One deeply divergent cluster is comprised of phytoplasmas together with *Acholeplasma* species and *Anaeroplasma* species forming a monophyletic group (Gundersen *et al.*, 1994; Lee *et al.*, 2000).

In the present study, a consensus phylogeny inferred by multiple sets of concatenated housekeeping proteins clearly delineated phylogenetic relationships among 41 representative walled bacteria and wall-less mollicutes, reinforcing the notion that phytoplasmas represent a distinct lineage evolving in parallel with mycoplasmas. The divergence between phytoplasmas and the other mollicutes is also indicated by the difference in the genetic code systems used by the phytoplasmas and the other mollicutes examined in this study.

The phylogenetic trees reconstructed from the five sets of concatenated housekeeping proteins imply similar degrees of conservation among the five sets of proteins analysed. The phylogenies inferred by these proteins, whose sequences are less conserved than those of 16S rRNA genes, clearly amplify the resolving power for delineating phylogenetic relationships among the prokaryotes studied. Thus, these housekeeping proteins should provide a more discriminating tool for phylogenetic analysis than the 16S rRNA gene.

ACKNOWLEDGEMENTS

Spiroplasma kunkelii genome sequence data were made available by Bruce Roe, ShaoPing Lin, HongGui Jia, HongMin Wu and Doris Kupfer (University of Oklahoma, Department of Chemistry and Bio-chemistry, Norman, OK 73019, USA) and Robert E. Davis (US Department of Agriculture-Agricultural Research Service, Molecular Plant Pathology Laboratory, Beltsville, MD 20705, USA). The *Spiroplasma kunkelii* genome sequencing project is funded by the US Department of Agriculture, Agricultural Research Service, project number 1275-22000-144-02. '*Candidatus* Phytoplasma asteris'-related strain AY-WB genome sequence data were made available by J. Zhang, L. Liefting, X. Bai, S. A. Miller, B. Kirkpatrick, J. Campbell, E. Goltsman, T. Walunas, N. Kyrpides and S. A. Hogenhout, 'Genome sequencing of phytoplasmas, pathogens of insects and plants: a consortium' funded by the US Department of Agriculture, Cooperative State Research, Education and Extension Service, award number 2002-35600-12752.

REFERENCES

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–3402.

Barnett, J. A. (2003). A history of research on yeasts 6: the main respiratory pathway. *Yeast* **20**, 1015–1044.

Besemer, J. & Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. *Nucleic Acids Res* **27**, 3911–3920.

Bové, J. M. (1997). Spiroplasmas: infectious agents of plants, arthropods and vertebrates. *Wien Klin Wochenschr* **109**, 604–612.

Brochier, C., Philippe, H. & Moreira, D. (2000). The evolutionary history of ribosomal protein RpS14: horizontal gene transfer at the heart of the ribosome. *Trends Genet* **16**, 529–533.

Brown, J. R. & Doolittle, W. F. (1997). Archaea and the prokaryote-to-eukaryote transition. *Microbiol Mol Biol Rev* **61**, 456–502.

Brown, J. R., Douady, C. J., Italia, M. J., Marshall, W. E. & Stanhope, M. J. (2001). Universal trees based on large combined protein sequence data sets. *Nat Genet* **28**, 281–285.

Canback, B., Andersson, S. G. & Kurland, C. G. (2002). The global phylogeny of glycolytic enzymes. *Proc Natl Acad Sci U S A* **99**, 6097–6102.

Castresana, J. (2000). Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**, 540–552.

Chattopadhyay, S. & Chakrabarti, J. (2003). Temporal changes in phosphoglycerate kinase coding sequences: a quantitative measure. *J Comput Biol* **10**, 83–93.

Choi, J. H., Jung, H. Y., Kim, H. S. & Cho, H. G. (2000). PHYLODRAW: a phylogenetic tree drawing system. *Bioinformatics* **16**, 1056–1058.

Ciccarese, S., Tommasi, S. & Vonghia, G. (1989). Cloning and cDNA sequence of the rat X-chromosome linked phosphoglycerate kinase. *Biochem Biophys Res Commun* **165**, 1337–1344.

Citti, C., Maréchal-Drouard, L., Saillard, C., Weil, J. H. & Bové, J. M. (1992). *Spiroplasma citri* UGG and UGA tryptophan codons: sequence of the two tryptophanyl-tRNAs and organization of the corresponding genes. *J Bacteriol* **174**, 6471–6478.

Daubin, V., Gouy, M. & Perriere, G. (2002). A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res* **12**, 1080–1090.

Davis, R. E., Worley, J. F., Whitcomb, R. F., Ishijima, T. & Steere, R. L. (1972). Helical filaments produced by a mycoplasma-like organism associated with corn stunt disease. *Science* **176**, 521–523.

den Blaauwen, T. & Driessen, A. J. (1996). Sec-dependent preprotein translocation in bacteria. *Arch Microbiol* **165**, 1–8.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science* **284**, 2124–2129.

Felsenstein, J. (1989). PHYLIP – phylogeny inference package (version 3.2). *Cladistics* **5**, 164–166.

Felsenstein, J. (2004). PHYLIP (phylogeny inference package), version 3.6. Department of Genome Sciences, University of Washington, Seattle, USA.

Felsenstein, J. & Churchill, G. A. (1996). A hidden Markov Model approach to variation among sites in rate of evolution. *Mol Biol Evol* **13**, 93–104.

Forterre, P. & Philippe, H. (1999). Where is the root of the universal tree of life? *Bioessays* **21**, 871–879.

Fothergill-Gilmore, L. A. (1986). The evolution of the glycolytic pathway. *Trends Biochem Sci* **11**, 47–51.

Fox, G. E., Wisotzkey, J. D. & Jurtshuk, P., Jr (1992). How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. *Int J Syst Bacteriol* **42**, 166–170.

Gasparich, G. E., Whitcomb, R. F., Dodge, D., French, F. E., Glass, J. & Williamson, D. L. (2004). The genus *Spiroplasma* and its non-helical descendants: phylogenetic classification, correlation with phenotype and roots of the *Mycoplasma mycoides* clade. *Int J Syst Evol Microbiol* **54**, 893–918.

Golding, G. B. & Gupta, R. S. (1995). Protein-based phylogenies support a chimeric origin for the eukaryotic genome. *Mol Biol Evol* **12**, 1–6.

Gundersen, D. E., Lee, I. M., Rehner, S. A., Davis, R. E. & Kingsbury, D. T. (1994). Phylogeny of mycoplasmalike organisms (phytoplasmas): a basis for their classification. *J Bacteriol* **176**, 5244–5254.

Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. (1998). Multiple sequence alignment with CLUSTAL X. *Trends Biochem Sci* **23**, 403–405.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* **8**, 275–282.

Klenk, H. & Zillig, W. (1994). DNA-dependent RNA polymerase subunit B as a tool for phylogenetic reconstructions: branching topology of the archaeal domain. *J Mol Evol* **38**, 420–432.

Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. (2001). MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.

Lee, I.-M., Davis, R. E. & Gundersen-Rindal, D. E. (2000). Phytoplasma: phytopathogenic mollicutes. *Annu Rev Microbiol* **54**, 221–255.

Makarova, K. S., Ponomarev, V. A. & Koonin, E. V. (2001). Two C or not two C: recurrent disruption of Zn-ribbons, gene duplication, lineage-specific gene loss, and horizontal gene transfer in evolution of bacterial ribosomal proteins. *Genome Biol* **2**, research 0033.1–0033.14.

Mollet, C., Drancourt, M. & Raoult, D. (1997). *rpoB* sequence analysis as a novel basis for bacterial identification. *Mol Microbiol* **26**, 1005–1011.

Olsen, G. J. & Woese, C. R. (1993). Ribosomal RNA: a key to phylogeny. *FASEB J* **7**, 113–123.

Oshima, K., Kakizawa, S., Nishigawa, H. & 8 other authors (2004). Reductive evolution suggested from the complete genome sequence of a plant-pathogenic phytoplasma. *Nat Genet* **36**, 27–29.

Philippe, H. & Forterre, P. (1999). The rooting of the universal tree of life is not reliable. *J Mol Evol* **49**, 509–523.

Pollack, J. D., Williams, M. V. & McElhaney, R. N. (1997). The comparative metabolism of the mollicutes (*Mycoplasmas*): the utility for taxonomic classification and the relationship of putative gene annotation and phylogeny to enzymatic function in the smallest free-living cells. *Crit Rev Microbiol* **23**, 269–354.

Razin, S., Yogev, D. & Naot, Y. (1998). Molecular biology and pathogenicity of mycoplasmas. *Microbiol Mol Biol Rev* **62**, 1094–1156.

Stackebrandt, E. & Goebel, B. M. (1994). Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Bacteriol* **44**, 846–849.

Stackebrandt, E., Rainey, F. A. & Ward-Rainey, N. L. (1997). Proposal for a new hierarchic classification system, *Actinobacteria* classis nov. *Int J Syst Bacteriol* **47**, 479–491.

Tatusov, R. L., Natale, D. A., Garkavtsev, I. V. & 7 other authors (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29**, 22–28.

Teeling, H., Lombardot, T., Bauer, M., Ludwig, W. & Glöckner, F. O. (2004). Evaluation of the phylogenetic position of the planctomycete '*Rhodopirellula baltica*' SH 1 by means of concatenated ribosomal protein sequences, DNA-directed RNA polymerase subunit sequences and whole genome trees. *Int J Syst Evol Microbiol* **54**, 791–801.

Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. (1997). The CLUSTAL_X Windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**, 4876–4882.

Tjalsma, H., Bolhuis, A., Jongbloed, J. D., Bron, S. & van Dijl, J. M. (2000). Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. *Microbiol Mol Biol Rev* **64**, 515–547.

Weisburg, W. G., Tully, J. G., Rose, D. L. & 9 other authors (1989). A phylogenetic analysis of the mycoplasmas: basis for their classification. *J Bacteriol* **171**, 6455–6467.

Woese, C. R. (1987). Bacterial evolution. *Microbiol Rev* **51**, 221–271.

Wolf, Y. I., Rogozin, I. B., Grishin, N. V., Tatusov, R. L. & Koonin, E. V. (2001). Genome trees constructed using five different approaches suggest new major bacterial clades. *BMC Evol Biol* **1**, 8.

Wolf, M., Müller, T., Dandekar, T. & Pollack, J. D. (2004). Phylogeny of *Firmicutes* with special reference to *Mycoplasma* (*Mollicutes*) as inferred from phosphoglycerate kinase amino acid sequence data. *Int J Syst Evol Microbiol* **54**, 871–875.